Archive of the Future: A Session for Information Gathering & Sharing

Cathleen Wharton & Mitch Black
Open Archive Development
Sun Microsystems, Inc.
Today's Session - Overview

• Presenters

• Logistical details (if needed, e.g. breaks, etc.)

• Primary Goals: Fact Finding & Information Sharing
  > Share what we're doing or considering for Open Archive
  > Generate discussion and Q&A around Focus areas: Many to choose from.
  > Extra Time? Grab bag: General “customer needs” questions.
What is the Sun Open Archive Team doing?
Open Archive to address multiple Market Sectors

• Digital Preservation
• Compliance & Governance
  > Business records
  > Regulatory processes
• As a Dynamic Active Data Store
  > Health records
  > Scientific community
  > Historical records
• Distinct from backup
Advanced Archive Services

- Ingestion
- Classification
- Policy
- Metadata
- User Metadata
- Index
- Search
- Global Namespace
- Open API
- Client User Interface
- Data Encoding
- Single Instance Storage

- Retention
- Deletion
- Permanent Deletion - Expunge
- Copy Management
- Remote Replication
- Integrity
- Scrub
- Audit Trail & Reporting (Data / Information)
- Verification (Fixity Check)
- Validation (Provenance)
- Migration
- Remastering (Physical Media)
- Transformation (Logical Format)
- Quarantine
- Identity Management
- Archive Extraction
- Archive Upgrade
- Archive Subsumption
- Archive Federation
- Scalability
- Archive Config & Reporting (Watermarks & Health)
Digital Preservation-focused Archive Services (Discussion)

- Ingestion
- Classification
- Policy
- Metadata
- User Metadata
- Index
- Search
- Global Namespace
- Open API
- Client User Interface
- Data Encoding
- Single Instance Storage
- Retention
- Deletion
- Permanent Deletion - Expunge
- Copy Management
- Remote Replication
- Integrity
- Scrub
- Audit Trail & Reporting (Data / Information)
- Verification (Fixity Check)
- Validation (Provenance)
- Migration
- Remastering (Physical Media)
- Transformation (Logical Format)
- Quarantine
- Identity Management
- Archive Extraction
- Archive Upgrade
- Archive Subsumption
- Archive Federation
- Scalability
- Archive Config & Reporting (Watermarks & Health)
Open Archive Deployment Flexibility
Open Archive Architecture

- **Archive System**
  - **Archive Services**
    - Index
    - Search
    - Policy
    - Classification
    - Retention
    - Integrity
    - Scrub
    - Audit Trail
  - **Customer Defined / Built Service**
  - **Copy Management**
  - **Deletion**

- **Existing User Store** (LDAP/Active Directory/NIS)
- **Metadata**
- **Back Up Store**
  - Disk
  - Objects (Data Content)
  - Tape

- **Interfaces**
  - **Administrative GUI**
  - **End-user GUI** (optional)
  - CIFS/SMB
  - NFS
  - FTP
  - XAM
  - Bulk Ingest
  - REST
  - Java API
  - XML
  - Fedora Commons ECM, ISVs, etc.

- **SOA Model**
  - **Choose Services**
  - **Package as appropriate for market**
  - **Replace easily**
  - **Agile Development**
Time to Focus
Which Specific Areas?

- Deployment
- Open Archive Features
- Security & Access
- Customization
- Interfaces
- Backing Store (short / long term)
- Metadata
- Retention / Copy Management

- Policy Definition & Exercising
- Critical areas of Performance
- User Experience
- Deduplication
- “Grab Bag” of Questions
What is your Preferred Deployment Model?

Appliance  Software  Cloud-based Service

Copyright 2008 Sun Microsystems, Inc. - Do Not Copy or Distribute Without Permission
What **Open Archive Features** are Most Critical?

- **Hardware**
  - General Purpose Hardware?
  - Tiers of Storage – SSD – Disk -- Tape?

- **Software**
  - Open Source Software and APIs?
  - Ability to Pick-&-Choose, Extend, Augment, and Customize Archive Services or the “Archive Experience”?

- **Management**
  - Solution Scalability – Horizontal and Vertical Scaling?
  - Management Simplicity?
  - Ability to Federate?
  - Ability to Collaborate?
  - Interoperability – Other Vendors for Software or Hardware?
What are your Key Security & Access Concerns?

• Archive Users
  > End-user access?
  > Physical location access?
  > Geo-political / cross-governmental, cross-citizenship access?
  > Federated access?

• Administrator Management
  > Authentication paradigms (e.g., LDAP, Active Directory)?
  > “Shrouded services” -- Service visibility?

• Data
  > Retention policies?
  > Secure delete?
  > Encrypted data?
  > Metadata – access or exposure?
What do you wish to Customize?

• **End-user role**
  > User interface branding to interact with the archive?
  > User interface capabilities exposed?
  > User-defined metadata submitted?

• **Administrator role (includes Chief Librarian / Compliance Office or other leader; in addition to storage/system administrator)**
  > Setup or integrate mechanisms for ingestion, display, federation?
  > Tailor data classification and policy needs to business environment?
  > Tailor archive workflow to match business workflow?

• **Archive Capabilities**
  > Ability to use select services (pick & choose)?
  > Ability to add in custom services unique to your mission?
  > Ability to extend select services and make these results visible?
  > Expand metadata capabilities?
What **Interfaces** are Most Needed / Most Useful?

- **Standards Based**: OAIS, XAM, Others?
- **Filesystem Based**: Direct FS access, CIFS, NFS, etc.
- **ECM Based**: FedoraCommons, etc.
- **GUI / BUI**: Wizard or other graphical methods
- **API**: Programmatic access
- **CLI**: Script and shell access
- **XML or other structured formats**
- **Also, Metadata input I/F?**
What Issues are Most Salient with Respect to your **Backing Store**?

- **Hardware migration**: to migrate to newer (cheaper / faster / higher capacity) technologies - Frequency?
- **Ability to utilize existing hardware** (disk / tape)
- **Expandability / Scalability**
- **Distributed archive**: across geographic / physical locations
- **Resource sharing**: storage, network bandwidth, drives
- **Resource utilization**: hotspots, problem areas, capacity planning
- **Flexibility in selection of hardware** (no vendor lock-in)
- **What other short-term hardware issues? Long-term?**
How should **Metadata** be Generated & Associated with Archived Objects?

**Automatic Generation / Association**
- POSIX “standard” file metadata (owner, size, dates, etc.)
- Content processing – Index, Keywords (cleartext)
- Standard-formats metadata extraction (TIFF, MP3, etc.)

**Manual or Site-Customized Generation / Association**
- Input by user (typing / select from lists) at ingestion time
- Generated & associated by onsite scripts with CLI / API
- Provided in structured format (XML) or schema
- Extracted from application-specific metadata inside data object
- Others?
How do you expect the Archive to be involved in Retention & Copy Mgmt?

- **Duration**: Any retention needs that are not “forever”?
- **Retention shortened?** Protection vs. Flexibility
- **Automated & Handled by policy**: Admin must define
- **User “Making an extra copy” vs. deduplication**?
- **Uncontrolled copies outside the archive.**
  - For retained archive objects, changed copies brought back in are whole new objects (user modified, previous version, etc.)
  - Backup is archival, not “traditional”: you keep more copies; copies can self-validate and inter-copy-correct.
How should Policies be Defined & Applied?

- Policy = (trigger or condition) + (action)
- Policy purpose: Automatically perform services
  > On archived objects
  > On components of the archive system
- Association: Policies assigned by classification only? How else?
- Some rudimentary defaults (templates) included with product
- Site specific policies must be defined... on site: Assumes a clear definition & consistency of desired policy.
- Timing: Exercise of policy ideally at “off-peak” hours.
- Expectations when policy activities cannot be completed in time allotted?
What are your most Critical Areas of Performance?

...And what would you ideally like to do about them???

• Ingestion

• Migration

• I/O (internal to archive, i.e. disk to disk, disk to tape)
  > Disk I/O contention
  > Tape Drive contention

• Network (EtherNet) bandwidth / SAN bandwidth

• CPU intensive operations (object processing)

• Object access (delay tolerance / search tolerance)
What is your **User Experience?**

- **End-users**
  - Submit data for inclusion in the archive
  - Add metadata to an information item in the archive
  - Search for, locate, and review something in the archive

- **Chief Compliance Officers / Chief Archival Officers**
  - Establish and maintain classifications for data in the archive
  - Establish and maintain policies for data in the archive
  - Establish and maintain ILM practice

- **Archive Administrators**
  - Configure appliance
    - To meet established policies
    - To integrate in standard APIs and existing ISV applications
  - Monitor appliance
  - Increase appliance capacity
  - Audit archive content
Is **Deduplication** a relatively High-Value item for you?

- **Quantify**: How much benefit would it give you?
- **Perceived risks**
  - Loss of one file is now loss of many
  - “When I make a copy, I have no more copies than I did.”
  - Deduplication and copy management are intertwined.
  - Compute intensive... but hashing / fingerprinting must be done anyway for validation & preservation.
- **Block level vs. file level** (granularity / layer / visibility)
Time Permitting

OR

For Follow-up Discussion

Grab Bag of Questions
Long-Lived Data...

Do you tend to keep long-lived data on disk or on tape?

Why?

Do you wish to change this?

Is most of your long-lived data backups (i.e. copies of things you have on disk, and for recovery purposes)?
Finding Archived Objects...

If you need to find a specific archived object, how do you do so?

Do you have any "wish-list" with regard to this ("I wish I could just....")?
Multiple Physical Locations...

Is it useful to you to have copies of the same data across multiple physical sites?

Do you do so today?

Why / Why not?
Access to data (interfaces)...

With regard to data management, do most of your client applications (or people using your data storage) share a common interface, or are they accessing disk and tape in many varied proprietary ways?

If it is the latter, how do you keep this from being an ongoing business cost?
Organizing data...

Do you subdivide your data/documents into "types" or "classes" in order to:

- put interrelated data into "buckets"?
- apply data management policies that vary depending on the types of data (i.e. saving, deletion, encryption, etc.)?
- make it easier to find specific data/documents when they are needed?

How do you subdivide data? Automatically? Through user-defined policy? through corporate defined policy? by corporate organization/function? Other?
Problems...

With regard to archived objects managed at your site, what is your biggest "headache"?
Next Steps...
- FIN -

Cathleen Wharton & Mitch Black

c.wharton @ sun.com  -  (303) 272-7543
mitch.black @ sun.com - (303) 272-7642